APPLICATIONS OF CLUSTER ANALYSIS TO SOME PROBLEMS OF INTEREST TO THE U. S. DEPARTMENT OF STATE

James Richard Lamping

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

APPLICATIONS OF CLUSTER ANALYSIS
TO SOME PROBLEMS OF INTEREST
TO THE U. S. DEPARTMENT OF STATE

by

James Richard Lamping

Thesis Advisor : B. Shubert

September 1973

Approved for public release; distribution unlimited.

Applications of Cluster Analysis to Some Problems of Interest to the U. S. Department of State

by

James Richard Lamping
Lieutenant Commander, United States Navy
B.S., University of Notre Dame, 1964

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1973

ABSTRACT

This thesis asserts that Cluster Analysis, or Numerical Taxonomy, has many potential applications in the field of international relations. It demonstrates two representative applications. Both examples treat the nations of the world as objects having measurable attributes, and both examples use selected attributes to produce a dendrogram (or hierarchical classification) of the nations of the world. In one example this dendrogram is used to objectively group the nations into blocs based on external economic ties. In the other example the dendrogram is used to highlight interactions among five attributes, ignoring the identity of individual nations, the same way a scatter plot highlights interactions between two variables.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DRAWINGS

I. BACKGROUND

A. CLUSTER ANALYSIS DEFINED

The subject of this thesis is a group of mathematical techniques known collectively as either Cluster Analysis or Numerical Taxonomy. The terms are equivalent. Their formal definition, paraphrased from Ref. 6, is actually a sequence of definitions, as follows

classification system - a set of subsets of a set of objects which conveys some information about the objects

taxonomy - the science of constructing classificatory systems

Cluster Analysis or Numerical Taxonomy - the science of constructing mathematical classificatory systems

In less formal terms, Cluster Analysis includes all mathematical methods of classifying objects into sets so as to represent complex data in a simpler way which will serve as a fruitful source of hypothesis.

Cluster Analysis is a two stage process. The first stage is to choose quantifiable attributes that describe the objects, and then use these attributes to measure the pair-wise dissimilarity among the objects. The second stage is to represent these dissimilarities by an appropriate classificatory system or display.

The input to Cluster Analysis is normally an $n \times m$ matrix of data, measurements of m attributes for each of n objects.

The output from Cluster Analysis is normally one of three displays:

A hierarchical classification, commonly called a tree diagram or dendrogram;

A partition of the objects into mutually exclusive sets, each set described by a "profile" or vector of m average attribute values;

A "clumping" of the objects into sets that may overlap, each set again described by a profile.

The value of these outputs is that they summarize the original data objectively and they tend to highlight subtle interactions in the original data, enabling a user to formulate reasonable hypotheses about these interactions.

B. PREVIOUS APPLICATIONS OF CLUSTER ANALYSIS

Cluster Analysis was developed in the eighteenth century by botanists and biologists attempting to inject more objectivity into their classifications of plant and animal specimens (the familiar phylum-genus-species scheme). Subsequently the same technique was used by geologists. Most recently, Cluster Analysis has found numerous applications in the social sciences, particularly in psychology. Reference 1 describes an application that is representative.

II. PROPOSED APPLICATIONS OF CLUSTER ANALYSIS IN THE STATE DEPARTMENT

The United States State Department is currently trying to revitalize its policy-making and resource-allocating functions, a la the Defense Department metamorphosis under Robert McNamara. This revitalization effort has been underway for eight years now. In that time there has been published a plethora (References 2 and 3 are representative) of "master plans" for the incorporation of Systems Analysis in the State Department.

This paper does not propose another master plan, but merely suggests that a single existing statistical analysis technique has useful applications within the State Department. The existing technique is Cluster Analysis, and the potential applications within the State Department are described and demonstrated in the pages that follow.

A. FIRST APPLICATION: TO HIGHLIGHT INTERACTIONS OF VARIABLES

1. General Description

In this application, Cluster Analysis highlights the interactions of several variables the same way a scatter plot would for two variables. It inputs an $n \times m$ matrix of data (n countries, each described by m variables) and outputs a dendrogram. The dendrogram itself says nothing about interactions among the variables. But it is a simple matter to select a clustering level (where k = the number of clusters)

and plot the distribution of the m variables within each cluster. Comparisons among these plots should bring out all significant interactions among the variables. In particular, it should highlight mutual interaction among three variables or even among four variables just as easily as it highlights a two-way interaction. This is a potential not shared by factor analysis and regression techniques.

2. Scenario for Demonstration

The United States Constitution lists Freedom of the Press, Freedom of Speech, and Freedom of Assembly as inalienable human rights. Although one might argue that these precise terms have been eclipsed by communications technology, most of the Western world would agree that "free and facile communication among the people" is an essential quality in a free and productive society. Having tentatively accepted this thesis, a sociologist or political scientist might well wish to dissect the concept of "free and facile communication among the people," to define it in quantifiable terms. Moreover, a policy planner in the State Department might well wish to go one step further: to use this quantitative definition in a comparative study of the countries of the world. Such comparisons are made every day with respect to Gross National Product, Life Expectancy, etc. Why not also tabulate a FFCAP (free and facile communication among the people) Index?

Assuming that the State Department considered it worthwhile to develop such an index, they would probably task a team of their sociologists to propose a list of measurable

factors that either contribute to or detract from "free and facile communication among the people." This team would certainly appreciate the practical advantages of building this list around statistics that had already been measured, and using the existing data to continually validate their theories against the real world.

Thus they would probably be faced, early on in their proceedings, with a large volume of existing data to be perused, or analyzed in a very general sense. At this point they could profit greatly from applying Cluster Analysis to highlight the interaction of variables.

3. Choice of Data

To demonstrate this application, the author has usurped the role of State Department sociologist and selected the following statistics as "measurable factors that either contribute to or detract from free and facile communication among the people":

Variable 1. Concentration of Population in Cities, 1965

Variable 2. Radios per 1000 Population, 1965

Variable 3. Students in Higher Education (Third Level) per One Million Population, 1965

Variable 4. Ethno-Linguistic Fractionalization

Variable 5. Press Freedom Index, 1965

See Appendix A for definitions of these variables.

It is readily admitted that this list is not as complete as it should be. In particular, "Literacy Rate" is conspicuous by its absence, and some measure of newspaper

circulation seems a necessary counterpart to Variable 2. The reason for such omissions was unavailability of data, an affliction that is widespread among independent researchers but not shared by insiders at the State Department.

The unavailability of data to this researcher imposed another artificiality on this demonstration besides the omission of some desirable variables. Table I displays values of the aforementioned variables for only 85 of the 136 nations in the world. It was necessary to delete the other nations because of excessive missing data.

4. Choice of Dissimilarity Coefficient

This section describes the process of converting the data in Table I to a matrix of Dissimilarity Coefficients.

The first decision point was to specify a formula for the Dissimilarity Coefficient (DC). The DC is a single real number specifying the amount of dissimilarity between Country A and Country B, obtained by somehow combining the five data points describing each country. There are many different formulas for transforming these ten data points to a single DC. Cormack presents a concise but comprehensive summary of all the common formulas in Table 1 of Ref. 4.

In the situation at hand it was decided to use a Euclidean Distance, standardized by range. That is,

$$DC(i,j) = \sum_{v=1}^{5} W_v (X_{iv} - X_{jv})^2$$

$$\text{where} \quad W_v = \frac{1}{\max_{i,j} (X_{iv} - X_{jv})^2}$$

TABLE I DATA INPUT TO HIGHLIGHTING DEMONSTRATION

COUNTRY NAME	COUNTRY CODE	VARIABLE 1 CONCEN. OF POPULATION IN CITIES	VARIABLE 2 RADIOS PER 1000 POPULATION	VARIABLE 3 STUDENTS IN HIGHER ED. PER MIL.POP	VARIABLE 4 ETHNO-LINGUISTIC FRACTION.	VARIABLE 5 PRESS FREEDOM INDEX
UNITED STATES	2	0.024	1233.5	28400.0	0.505	2.72
CANADA	20	0.044	518.9	16510.0	0.755	2.78
CUBA	40	0.031	180.9	4C00.0	0.033	-3.02
DOMINICAN REPUBLIC	42	0.052	40.0	1820.0	0.037	1.16
MEXICO	70	0.014	193.0	3120.0	0.305	1.46
EL SALVADOR	92	0.025	135.2	1240.0	0.166	2.26
COSTA RICA	94	0.031	90.7	5040.0	0.072	2.68
COLOMBIA	100	0.025	118.0	214C.0	0.060	2.21
VENEZUELA	101	0.055	190.3	5370.0	0.107	2.54
EQUADOR	130	0.025	100.3	2990.0	0.534	2.12
PERU	135	0.030	185.9	4230.0	0.590	2.76
BRAZIL	140	0.012	95.2	1890.0	0.071	1.25
BOLIVIA	145	0.024	142.0	3630.0	0.678	2.39
CHILE	155	0.076	240.0	5080.0	0.140	1.19
ARGENTINA	160	0.037	295.0	10890.0	0.307	0.92
URUGUAY	165	0.210	339.3	6100.0	0.198	2.61
UNITED KINGDOM	200	0.046	296.6	4857.1	0.325	2.37
IRELAND	205	0.047	212.3	7590.0	0.045	2.37
NETHERLANDS	210	0.023	251.6	12090.0	0.102	3.02
BELGIUM	211	0.039	319.7	8050.0	0.551	2.53
FRANCE	220	0.033	313.5	1042C.0	0.261	1.92
SWITZERLAND	225	0.033	278.2	5540.0	0.504	3.06
SPAIN	230	0.020	144.0	4170.0	0.436	-1.00
PORTUGAL	235	0.012	127.5	3530.0	0.006	-1.43
WEST GERMANY	255	0.014	440.4	6320.0	0.026	2.43
EAST GERMANY	265	0.020	336.9	4660.0	0.017	-3.20
POLAND	290	0.015	179.3	8C00.0	0.028	-2.54
AUSTRIA	305	0.064	296.9	6810.0	0.126	2.10
HUNGARY	310	0.045	244.8	5C3C.0	0.098	-1.58
CZECHOSLOVAKIA	315	0.015	263.2	10C10.0	0.490	-2.51
ITALY	325	0.018	207.9	5830.0	0.038	1.98
ALBANIA	339	0.030	70.0	6840.0	0.093	-3.51
YUGOSLAVIA	345	0.011	153.8	9480.0	0.754	0.08
GREECE	350	0.057	105.7	6260.0	0.099	1.37
CYPRUS	352	0.053	218.9	133C.0	0.349	1.96
BULGARIA	355	0.020	250.6	12200.0	0.220	-2.71
RUMANIA	360	0.013	146.6	686C.0	0.252	-3.20
SOVIET UNION	365	0.003	320.0	16740.0	0.666	-3.08
FINLAND	375	0.035	334.1	8410.0	0.159	2.72
SWEDEN	380	0.024	381.9	9230.0	0.083	2.83
NORWAY	385	0.030	292.5	5250.0	0.039	3.06

TABLE I , CONTINUED

NAME	CODE	VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5
DENMARK	390	0.033	333.5	10050.0	0.049	2.65
SENEGAL	433	0.029	58.8	790.0	0.723	-1.99
UPPER VOLTA	439	0.003	10.5	6.0	0.678	-3.09
GHANA	452	0.007	73.6	550.0	0.706	0.34
CAMEROON	471	0.007	22.0	300.0	0.892	-2.42
NIGERIA	475	0.003	10.6	160.0	0.869	0.45
CHAD	483	0.002	8.0	230.0	0.826	-2.72
UGANDA	500	0.0	26.5	160.0	0.899	0.77
KENYA	501	0.002	37.4	300.0	0.833	1.20
TANZANIA	510	0.001	10.9	50.0	0.926	0.87
ETHIOPIA	530	0.001	14.6	100.0	0.694	-3.10
ZAMBIA	551	0.024	12.0	170.0	0.818	1.05
RHODESIA	552	0.011	12.2	200.0	0.544	1.16
MALAWI	553	0.001	20.0	110.0	0.620	0.52
SOUTH AFRICA	560	0.021	145.5	3290.0	0.877	1.07
MOROCCO	600	0.020	52.5	780.0	0.534	1.00
ALGERIA	615	0.017	129.0	680.0	0.435	-3.26
TUNISIA	616	0.039	67.0	1410.0	0.158	-0.67
IRAN	630	0.023	67.0	1050.0	0.756	-1.03
TURKEY	640	0.008	78.6	2930.0	0.255	1.66
IRAQ	645	0.044	349.0	3470.0	0.362	-1.36
UNITED ARAB REPUBLIC	651	0.026	54.5	5980.0	0.044	-2.32
SYRIA	652	0.033	329.2	6040.0	0.223	-2.00
LEBANON	660	0.090	120.6	8440.0	0.135	1.18
JORDAN	663	0.099	136.1	1620.0	0.047	-0.52
ISRAEL	666	0.056	290.0	14000.0	0.199	1.75
AFGHANISTAN	700	0.002	13.3	220.0	0.658	-1.30
CHINA	710	0.001	11.8	1220.0	0.118	-3.16
TAIWAN	713	0.030	102.2	6870.0	0.350	0.61
SOUTH KOREA	732	0.029	69.1	5000.0	0.0	0.42
JAPAN	740	0.025	208.5	11400.0	0.015	2.44
INDIA	750	0.001	11.0	2840.0	0.886	0.98
PAKISTAN	770	0.003	5.4	2670.0	0.645	-0.02
BURMA	775	0.002	10.7	850.0	0.475	0.38
CEYLON	780	0.004	39.0	1290.0	0.467	1.14
THAILAND	800	0.004	54.2	1660.0	0.664	0.70
CAMBODIA	811	0.005	157.8	1200.0	0.297	-1.15
LAOS	812	0.010	25.5	70.0	0.600	-0.47
SOUTH VIETNAM	817	0.071	63.6	1680.0	0.190	-0.45
MALAYSIA	820	0.014	44.7	1366.6	0.716	1.66
PHILIPPINES	840	0.005	39.2	14410.0	0.745	2.66
INDONESIA	850	0.003	7.8	950.0	0.764	-0.40
AUSTRALIA	900	0.104	222.2	11590.0	0.316	2.53
NEW ZEALAND	920	0.064	243.9	21000.0	0.373	2.24

Euclidean Distance was preferred to others simply because of its geometric, intuitive appeal.

On the other hand, a more elaborate rationale went into the decision to standardize by range. First of all it was decided that some type of standardizing (scaling) would be appropriate. Most of the literature argues convincingly that scaling is inappropriate when the difference in scale between two variables may be intrinsic; but no such intrinsic differences seemed likely in the five variables used here. Moreover, using unstandardized Euclidean Distance in this situation would clearly result in the DC being driven by Variables 2 and 3 while Variables 1 and 4 would be virtually ignored, and there is no a priori reason to intentionally emphasize one variable over another in this application.

Having decided to use some type of scaling, there were many types to choose from, namely scaling by standard deviation, scaling by range, and scaling by some other heterogeneity measure (see page 326 of Ref. 4 for a comparative discussion). Since the data distributions were mixed there was no compelling theoretical reason for choosing one scaling method over another. Eventually, scaling by range was selected for its simplicity. It would be interesting to see if scaling by standard deviation would significantly change the end result (dendrogram) from that obtained here; but this was not done.

Using Euclidean Distance standardized by range, the 425 data points in Table I were transformed into a matrix of 3570 Dissimilarity Coefficients.

5. Choice of Algorithm

There are several methods of proceeding from the matrix of Dissimilarity Coefficients (DC) to a partition (or a dendrogram of partitions) of the countries. Choosing one was the next decision point in this demonstration. All methods in general use fall into one of three categories:

a. Agglomerative Algorithms - a series of successive fusions of the 85 countries into groups.

b. Divisive Algorithms - a partitioning of the complete set of countries successively into finer partitions.

c. Reallocative Algorithms - successive reallocation of individual countries between the sets of some initial partition.

It was first decided that a reallocative algorithm would be inappropriate because it requires an initial partition, and there was no a priori evidence to suggest what that partition should be. Between the two remaining alternatives, theoretical considerations did not yield a preference: in nearly every case, agglomerative and divisive algorithms produce identical dendrograms. The agglomerative algorithm was selected because its details have been more thoroughly documented in the literature.

Within the family of agglomerative algorithms there are at least eight documented alternative "sorting strategies" or formulas for determining the DC between cluster (k) and cluster (ij), using the DC between cluster (k) and cluster (i) and the DC between cluster (k) and cluster (j). If the matrix of Dissimilarity Coefficients contains natural and compelling clusters, each having strong internal cohesion and strong

external isolation, then the choice of sorting strategy is not a critical one. But if natural and compelling clusters are not present, different sorting strategies can produce markedly different dendrograms. The eight common sorting strategies are explained in Chapter 3 of Ref. 4. Of those eight, the Complete Linkage-Furthest Neighbor sorting strategy and the Single Linkage-Nearest Neighbor sorting strategy represent the extremes. The others may be thought of as compromises between these two. The Complete Linkage-Furthest Neighbor sorting strategy can be expressed mathematically as

$$DC(k,ij) = \max(DC(k,i), DC(k,j))$$

It produces compact clusters having high internal cohesion; but it may sacrifice external isolation when natural and compelling clusters are not intrinsic in the data. At the other extreme, the Single Linkage-Nearest Neighbor sorting strategy can be expressed mathematically as

$$DC(k,ij) = \min(DC(k,i), DC(k,j))$$

It tends to produce chains of objects in addition to, or instead of, compact clusters, especially when natural and compelling clusters are not intrinsic in the data. In some applications this tendency is desirable.

For the demonstration at hand compact clusters were considered far more desirable than chains, and the Complete Linkage-Furthest Neighbor sorting strategy was selected. It would be interesting to see if one of the compromise sorting

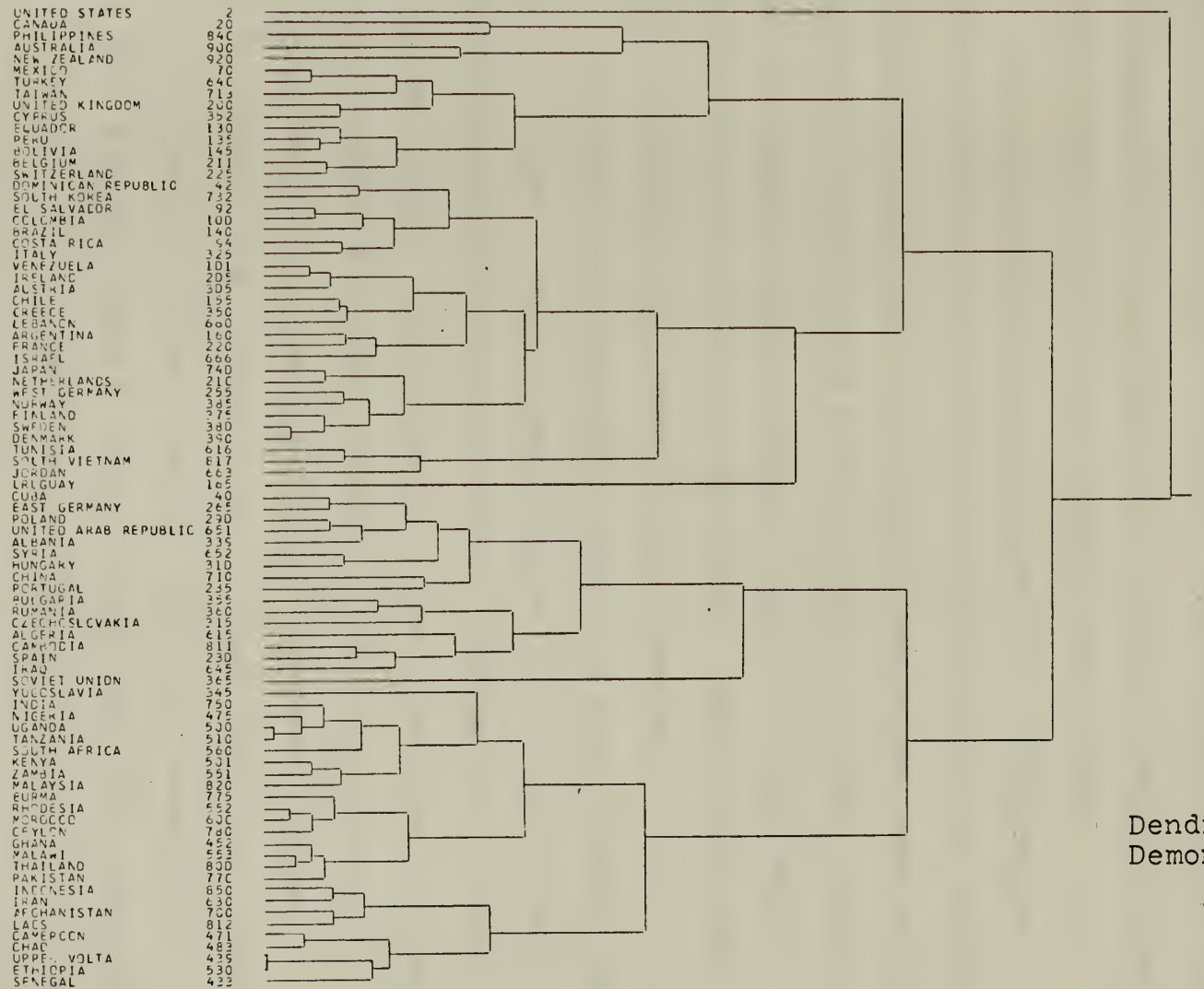
strategies, such as Group Average, would significantly change the end result (dendrogram) from that obtained here; but this was not done.

The dendrogram in Drawing 1 was obtained using the Complete Linkage-Furthest Neighbor sorting strategy in an agglomerative algorithm. The computer program is listed at the end of this thesis for information. It should be noted that the matrix of Dissimilarity Coefficients were standardized to the (0.0, 100.0) interval, using scaling by range, before they were input to the clustering algorithm. But such standardizing was made for computational convenience only. Its single effect was a monotonic transformation of the numerical scale across the top of the dendrogram. The shape of the dendrogram was unaffected.

6. From Dendrogram to Cluster Profiles

Having obtained the dendrogram in Drawing 1, and recalling that the purpose here was to highlight the interactions among variables, it remained only to select a level of clustering, identify the partition of countries there, plot the distributions of variables within each cluster, and compare these plots. But several iterations of this process were required before the interactions among variables began to appear.

The first attempt was at level k = 3. Here the United States appeared alone in one cluster, and the other two clusters contained 41 countries and 43 countries respectively. Apparently the United States was alone because



Drawing 1

Dendrogram in Highlighting Demonstration

of its extreme values in variables 2 and 3. At this point the United States was evaluated as an outlier and was not included in subsequent comparisons. Within each of the other two clusters, the mean and standard deviation of each of the five variables were computed. After a short perusal of these 20 statistics it became apparent that no interactions among variables were highlighted at this level. Within four of the five variables, the two means were displaced from each other by less than the sum of their standard deviations.

For the second iteration, level k = 7 was chosen. Again the pair-wise displacements between means were compared to standard deviations. The standard deviations were definitely smaller here than they had been at the k = 3 level: within cluster homogeneity had improved. Between cluster heterogeneity had improved to a lesser extent: many of the means were well separated but several others were not.

For the third iteration, level k = 8 was chosen. Here there were three clusters containing only one country each. All three were dismissed as outliers, leaving five clusters for further study. Within each of these five clusters, the mean and standard deviation of each of the five variables were computed, using a straightforward computer program. These statistics are listed in Table II and displayed graphically in Table III. After a relatively brief perusal of Table III, several possible interactions among the variables came to mind. Then a quick double-check of Table II confirmed

that four of those possible interactions were probable interactions. These probable interactions are listed in Table IV.

None of these "probable interactions" was verified mathematically. The first one would have been relatively easy to check out, by computing 10 correlation coefficients. But the others would have required considerably more ingenuity. Since the purpose here was to demonstrate a new application of cluster analysis rather than to deduce substantive results, mathematical verification was considered beyond the scope of this thesis.

However, there was a further step, within the scope of this thesis, that might have been pursued but was not. It would have been logical to proceed next to another cluster level (perhaps k = 13) and again look for interactions. Such reiterations might well confirm or refine the interactions already deduced, and highlight additional interactions as well.

B. SECOND APPLICATION: TO CLASSIFY COUNTRIES OBJECTIVELY

1. General Description

In this application, the user presumes to understand the variables used, and the interactions among these variables, at least on a superficial level. The purpose here is not to research the variables, but rather to objectively classify the countries. The previous application (to highlight interactions among variables) produced a dendrogram only as an intermediate step before producing "cluster profiles" as the final product. But the current application seeks only the dendrogram itself,

TABLE II CLUSTER PROFILES OUTPUT FROM HIGHLIGHTING DEMONSTRATION

CCUNTRY NAME	COUNTRY CODE	VARIABLE 1 CONCEN. OF POPULATION IN CITIES	VARIABLE 2 RADIOS PER 1000 POPULATION	VARIABLE 3 STUDENTS IN HIGHER ED. PER MIL.POP	VARIABLE 4 ETHNO-LINGUISTIC FRACTION.	VARIABLE 5 PRESS FREEDOM INDEX
UNITED STATES	2	0.024	1233.5	28400.0	0.505	2.72
CLUSTER WAS SINGULAR						
CANADA	20	0.044	518.9	16510.0	0.755	2.78
PHILIPPINES	840	0.005	39.2	14410.0	0.745	2.66
AUSTRALIA	900	0.104	222.2	11590.0	0.316	2.53
NEW ZEALAND	920	0.064	243.9	21000.0	0.373	2.24
CLUSTER PROFILE: AVGS:		0.05	256.	15878.	0.55	2.6
SDEVS:		0.04	198.	3966.	0.24	0.2
MEXICO	70	0.014	193.0	3120.0	0.305	1.46
TURKEY	640	0.008	78.6	2930.0	0.255	1.66
TAIWAN	713	0.030	102.2	6870.0	0.350	0.61
UNITED KINGDOM	200	0.046	296.6	4857.1	0.325	2.37
CYPRUS	352	0.053	218.9	1330.0	0.349	1.96
EQUADOR	130	0.025	100.3	2990.0	0.534	2.12
PERU	135	0.030	185.9	4230.0	0.590	2.76
BOLIVIA	145	0.024	142.0	3630.0	0.678	2.39
BELGIUM	211	0.039	319.7	8050.0	0.551	2.53
SWITZERLAND	225	0.033	278.2	5540.0	0.504	3.06
CLUSTER PROFILE: AVGS:		0.03	192.	4355.	0.44	2.1
SDEVS:		0.01	86.	2021.	0.14	0.7

TABLE II, CONTINUED

NAME	CODE	VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5
DOMINICAN REPUBLIC	42	0.052	40.0	1820.0	0.037	1.16
SOUTH KOREA	732	0.029	69.1	5000.0	0.0	0.42
EL SALVADOR	92	0.025	135.2	1240.0	0.166	2.26
COLOMBIA	100	0.025	118.0	2140.0	0.060	2.21
BRAZIL	140	0.012	95.2	1890.0	0.071	1.25
COSTA RICA	94	0.031	90.7	5040.0	0.072	2.68
ITALY	325	0.018	207.9	5830.0	0.038	1.98
VENEZUELA	101	0.055	190.3	5370.0	0.107	2.54
IRELAND	205	0.047	212.3	7590.0	0.045	2.37
AUSTRIA	305	0.064	296.9	6810.0	0.126	2.10
CHILE	155	0.076	240.0	5080.0	0.140	1.19
GREECE	350	0.057	105.7	6260.0	0.099	1.37
LEBANON	660	0.090	120.6	8440.0	0.135	1.18
ARGENTINA	160	0.037	295.0	10890.0	0.307	0.92
FRANCE	220	0.033	313.5	10420.0	0.261	1.92
ISRAEL	666	0.056	290.0	14000.0	0.199	1.75
JAPAN	740	0.025	203.5	11400.0	0.015	2.44
NETHERLANDS	210	0.023	251.6	12090.0	0.102	3.02
WEST GERMANY	255	0.014	440.4	6320.0	0.026	2.43
NORWAY	385	0.030	292.5	5250.0	0.039	3.06
FINLAND	375	0.035	334.1	8410.0	0.159	2.72
SWEDEN	380	0.024	381.9	9230.0	0.083	2.83
DENMARK	390	0.033	333.5	10050.0	0.049	2.65
TUNISIA	616	0.039	67.0	1410.0	0.158	-0.67
SOUTH VIETNAM	817	0.071	63.6	1680.0	0.190	-0.45
JORDAN	663	0.099	136.1	1620.0	0.047	-0.52
CLUSTER PROFILE: AVGS:		0.04	205.	6357.	0.11	1.7
SDEVS:		0.02	111.	3725.	0.08	1.1
URUGUAY	165	0.210	339.3	6100.0	0.198	2.61

CLUSTER WAS SINGULAR

TABLE II, CONTINUED

NAME	CODE	VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5
CUBA	40	0.031	180.9	4000.0	0.038	-3.02
EAST GERMANY	265	0.020	336.9	4660.0	0.017	-3.20
POLAND	290	0.015	179.3	8000.0	0.028	-2.54
UNITED ARAB REPUBLIC	651	0.026	54.5	5980.0	0.044	-2.32
ALBANIA	339	0.030	70.0	6840.0	0.093	-3.51
SYRIA	652	0.033	329.2	6040.0	0.223	-2.00
HUNGARY	310	0.045	244.8	5030.0	0.098	-1.58
CHINA	710	0.001	11.8	1220.0	0.118	-3.16
PORTUGAL	235	0.012	127.5	3530.0	0.006	-1.43
BULGARIA	355	0.020	250.6	12200.0	0.220	-2.71
RUMANIA	360	0.013	146.6	6860.0	0.252	-3.20
CZECHOSLOVAKIA	315	0.015	263.2	10010.0	0.490	-2.51
ALGERIA	615	0.017	129.0	680.0	0.435	-3.26
CAMBODIA	811	0.005	157.8	1200.0	0.297	-1.15
SPAIN	230	0.020	144.0	4170.0	0.436	-1.00
IRAQ	645	0.044	349.0	3470.0	0.362	-1.36
CLUSTER PROFILE: AVGS:		0.02	186.	5243.	0.20	-2.4
SDEVS:		0.01	102.	3137.	0.17	0.8
SOVIET UNION	365	0.003	320.0	16740.0	0.666	-3.08

CLUSTER WAS SINGULAR

TABLE II, CONTINUED

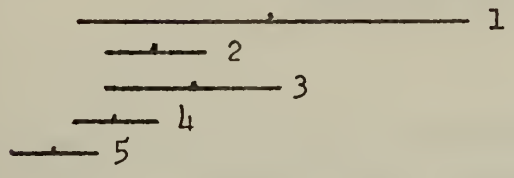
NAME	CODE	VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5
YUGOSLAVIA	345	0.011	153.8	9480.0	0.754	0.08
INDIA	750	0.001	11.0	284C.0	0.886	0.98
NIGERIA	475	0.003	10.6	160.0	0.869	0.45
UGANDA	500	0.0	26.5	160.0	0.899	0.77
TANZANIA	510	0.001	10.9	50.0	0.926	0.87
SOUTH AFRICA	560	0.021	145.5	329C.0	0.877	1.07
KENYA	501	0.002	37.4	300.0	0.833	1.20
ZAMBIA	551	0.024	12.0	170.0	0.818	1.05
MALAYSIA	820	0.014	44.7	1366.6	0.716	1.66
BURMA	775	0.002	10.7	850.0	0.475	0.38
RHODESIA	552	0.011	12.2	200.0	0.544	1.16
MOROCCO	600	0.020	52.5	780.0	0.534	1.00
CEYLON	780	0.004	39.0	1290.0	0.467	1.14
GHANA	452	0.007	73.6	550.0	0.706	0.34
MALAWI	553	0.001	20.0	110.0	0.620	0.52
THAILAND	800	0.004	54.2	1660.0	0.664	0.70
PAKISTAN	770	0.003	5.4	2670.0	0.645	-0.02
INDONESIA	850	0.003	7.8	950.0	0.764	-0.40
IRAN	630	0.023	67.0	1050.0	0.756	-1.03
AFGHANISTAN	700	0.002	13.3	220.0	0.658	-1.30
LACS	812	0.010	25.5	70.0	0.600	-0.47
CAMEROON	471	0.007	22.0	300.0	0.892	-2.42
CHAD	483	0.002	8.0	230.0	0.826	-2.72
UPPER VOLTA	439	0.003	10.5	6.0	0.678	-3.09
ETHIOPIA	530	0.001	14.6	100.0	0.694	-3.10
SENEGAL	433	0.029	58.8	790.0	0.723	-1.99
CLUSTER PROFILE: AVGS:		0.01	-36.	-114C.	0.72	-0.1
SDEVS:		0.01	39.	1929.	0.13	1.5
PROFILE FOR AGGREGATE OF 85 COUNTRIES: AVGS:		0.03	-166.	-5142.	0.38	-0.4
SDEVS:		0.03	169.	5198.	0.29	2.0

Table III

CLUSTER PROFILES, OUTPUT FROM HIGHLIGHTING DEMONSTRATION, GRAPHED

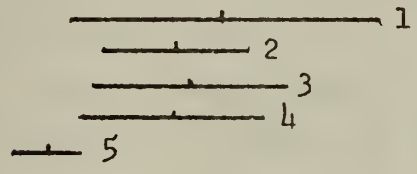
Variable 1. Concen. of Population in Cities

min=0.0 .05 .10 .15 0.21=max



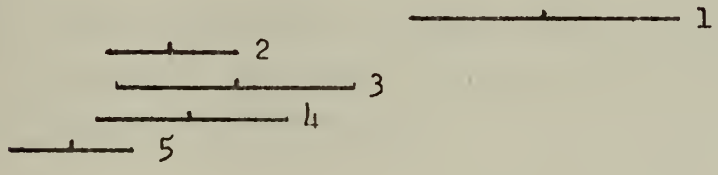
Variable 2. Radios per 1000 Population

min=5.4 200 400 600 800 1000 1233.5=max



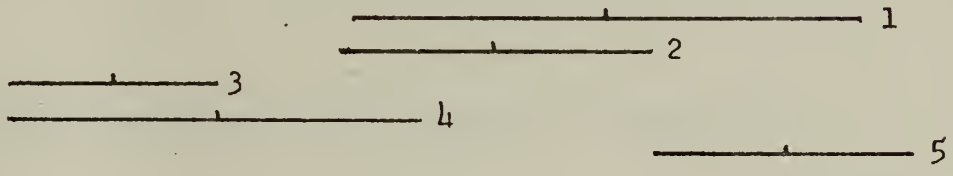
Variable 3. Students in Higher Ed. per Mil. Pop.

min=6.0 8000 16000 24000 28400=max



Variable 4. Ethno-Linguistic Fraction.

min=0.0 .2 .4 .6 .8 0.926=max



Variable 5. Press Freedom Index

min=-3.51 -2 0 2 3.06=max

ANALYSIS

TABLE IV

RESULTS OF HIGHLIGHTING DEMONSTRATION:
PROBABLE INTERACTIONS AMONG VARIABLES
DEDUCED AT CLUSTER LEVEL k = 8

1. There appear to be no significant pair-wise correlations (either positive or negative) among the five variables.

2. A high value in Variable 3 tends to be accompanied by a high value in Variable 5. But the inverse and converse are not true (i.e., a high value in Variable 5 does not imply a high value in Variable 3, and a low value in Variable 3 does not imply a low value in Variable 5).

3. A very high value in Variable 4 tends to be accompanied by a low value in Variables 1, 2, and 3.

4. The combination of high value in Variable 1 and a low value in Variable 4 tends to be accompanied by a high value in Variable 5.

****

For ready reference, the variable names are:

Variable 1. Concentration of Population in Cities, 1965

Variable 2. Radios per 1000 Population, 1965

Variable 3. Students in Higher Education (Third Level) per One Million Population, 1965

Variable 4. Ethno-Linguistic Fractionalization

Variable 5. Press Freedom Index, 1965

to objectively confirm, or perhaps modify, the user's previous, subjective, classifications.

2. Scenario for Demonstration

People commonly think about the countries of the world as members of clusters. They use labels like "the Western World", "the Communist Bloc", "the Have's" and "the Have-not's" every day, and they frequently hear more esoteric terms like "tri-polar world", "five-polar world" and "spheres of influence", all of which have classificatory overtones.

No doubt such classifications are convenient and useful; but as they exist now, many are also subjective and confusing. When two speakers discuss the behavior of "the Communist Bloc" without first enumerating the members of that bloc, they may disagree violently until they discover that one of them includes Cuba and Chile in his definition but exludes Yugoslavia, while the other has done the reverse. If our classifications of countries are useful but subjective, it would seem desirable to make them more objective.

Imagine that a political scientist in the State Department wished to inject some objectivity into the terms "Western World", "Communist Bloc", "Soviet Bloc", etc. His first step would probably be to identify the several theories (form of government, internal economic system, external political ties, external economic ties, etc.) that are commonly used to define the terms in question. Then he would probably select one of these theories for quantification and search out measurable factors (preferably statistics that had already been measured) with which to express it.

For example, imagine that he selected the theory that external economic ties are the prime mover in the concept of bloc membership. Then his search for measurable factors would certainly lead to statistics such as level of foreign aid received from every other country, value of imports received from every other country, and value of exports sent to every other country, each of these statistics prorated against the host country's GNP and/or population.

The final step for our State Department researcher would be to combine these measurable factors mathematically so as to output a bloc membership label for each country of the world; that is, he would write a "factors-to-bloc transformation". If he were not acquainted with Cluster Analysis, he might well try to write a single function of the form

$$\text{bloc membership} = F(\text{aid}_{US}, \text{aid}_{USSR}, \text{aid}_{UK}, \ldots$$
$$\text{trade}_{US}, \text{trade}_{USSR}, \text{trade}_{JAPAN}, \text{trade}_{COM.MKT.},$$
$$\text{GNP, population, etc.})$$

where "bloc membership" is a discrete variable which can take on three or perhaps five predetermined values. But such a function would probably be crippled by two weaknesses: excessive complexity and theoretical inadequacy. The reader can certainly visualize how complicated such a function would have to be in order to have broad applicability. Moreover, no matter how complex the function, it would necessarily ignore an obvious fact about blocs of countries: two countries can

be closely bound in a bloc not by economic dependency on each other but by their simultaneous economic dependency on an intermediate country.

Cluster Analysis has far more potential as a "factors-to-bloc" transformation. It does not share the dual weaknesses of the functional transformation. First of all, the ability to group two countries together through an intermediary is intrinsic to every clustering algorithm (so long as the Complete Link-Furthest Neighbor sorting strategy is not used). And secondly, Cluster Analysis requires that the user define only a transformation from measurable factors to a pair-wise Dissimilarity Coefficient rather than a transformation from measurable factors to bloc membership. Surely the former should be less complex than the latter.

3. Choice of Data

To demonstrate this application, the author again usurped the role of State Department political scientist and selected the following statistics as measurable factors with which to objectively classify countries into blocs:

Variable 1. Gross National Product per Capita, 1965

Variable 2. Trade as percentage of Gross National Product, 1965

Variable 3. Soviet Aid per Capita, 1954 - 1965

Variable 4. U.S. Economic Aid per Capita, 1958 - 1965

Values of these variables for each of 85 countries are listed in Table V.

TABLE V - DATA INPUT TO OBJECTIVE CLASSIFICATION DEMONSTRATION

COUNTRY NAME	COUNTRY CODE	VARIABLE 1 GNP PER CAPITA	VARIABLE 2 TRADE AS % OF GNP	VARIABLE 3 USSR AID PER CAPITA	VARIABLE 4 U.S. AID PER CAPITA
UNITED STATES	2	3574.5	7.3	0.0	0.0
CANADA	20	2472.6	34.8	0.0	0.0
CUBA	40	393.1	51.7	16.07	2.32
DOMINICAN REPUBLIC	42	265.3	23.3	0.0	51.92
MEXICO	70	455.2	13.9	0.0	15.25
EL SALVADOR	92	271.2	49.1	0.0	29.95
COSTA RICA	94	413.1	49.0	0.0	61.20
COLOMBIA	100	282.4	19.5	0.0	32.07
VENEZUELA	101	881.9	53.4	0.0	31.56
ECUADOR	130	215.6	28.7	0.0	32.36
PERU	135	367.1	32.5	0.0	28.78
BRAZIL	140	267.2	12.3	0.0	17.95
BOLIVIA	145	163.6	37.9	0.0	76.63
CHILE	155	565.2	26.7	0.0	96.39
ARGENTINA	160	769.7	15.7	5.14	16.19
URUGUAY	165	572.7	22.0	0.0	25.19
UNITED KINGDOM	200	1818.1	29.1	0.0	4.62
IRELAND	205	979.5	59.2	0.0	0.0
NETHERLANDS	210	1554.3	72.6	0.0	0.24
BELGIUM	211	1803.8	74.7	0.0	0.10
FRANCE	220	1924.0	21.7	0.0	0.18
SWITZERLAND	225	2332.9	48.0	0.0	0.0
SPAIN	230	561.4	22.5	0.0	23.17
PORTUGAL	235	405.6	40.2	0.0	12.97
WEST GERMANY	255	1900.9	31.5	0.0	1.03
EAST GERMANY	265	1260.0	25.9	4.46	0.0
POLAND	290	977.9	20.5	9.48	2.43
AUSTRIA	305	1286.8	39.6	0.0	9.59
HUNGARY	310	1093.8	39.9	11.97	0.0
CZECHOSLOVAKIA	315	1560.8	47.1	3.90	0.0
ITALY	325	1104.1	25.5	0.0	9.99
ALBANIA	339	365.7	24.0	15.76	0.0
YUGOSLAVIA	345	451.1	27.0	1.19	47.00
GREECE	350	687.3	24.9	9.82	41.89
CYPRUS	352	695.3	53.0	0.0	32.15
BULGARIA	355	829.3	67.5	63.39	0.0
RUMANIA	360	777.8	32.2	11.46	0.0
SOVIET UNION	365	1357.3	9.8	0.0	0.0
FINLAND	375	1749.1	38.1	0.0	3.25
SWEDEN	380	2549.0	42.3	0.0	0.0
NORWAY	385	1890.4	51.9	0.0	13.43

TABLE V, CONTINUED

NAME	CODE	VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4
DENMARK	390	2120.2	51.0	0.0	0.0
SENEGAL	433	194.8	42.9	2.01	4.73
UPPER VOLTA	439	52.9	24.2	0.0	1.09
GHANA	452	285.1	33.5	11.50	21.24
CAMEROON	471	128.1	43.4	1.53	4.74
NIGERIA	475	84.4	31.4	0.0	2.74
CHAD	483	71.7	24.5	0.0	1.30
UGANDA	500	87.1	44.5	2.12	2.25
KENYA	501	90.3	46.6	4.70	3.31
TANZANIA	510	71.4	42.1	0.0	4.21
ETHIOPIA	530	45.1	26.3	4.51	5.11
ZAMBIA	551	213.5	108.2	0.0	0.75
RHODESIA	552	239.7	79.4	0.0	0.49
MALAWI	553	47.0	53.0	0.0	2.06
SOUTH AFRICA	560	610.7	35.9	0.0	0.34
MOROCCO	600	195.5	33.8	0.0	34.04
ALGERIA	615	221.5	66.5	19.37	13.56
TUNISIA	616	214.1	38.6	6.34	97.49
IRAN	630	251.4	36.7	14.09	18.29
TURKEY	640	282.3	11.8	6.76	44.80
IRAQ	645	231.1	69.8	22.27	4.51
UNITED ARAB REPUBLIC	651	158.8	32.7	34.16	31.92
SYRIA	652	212.3	33.8	28.30	15.45
LEBANON	660	436.6	51.6	0.0	16.80
JORDAN	663	255.6	36.0	0.0	209.46
ISRAEL	666	1422.2	33.5	0.0	207.49
AFGHANISTAN	700	83.1	16.1	36.68	14.53
CHINA	710	108.6	6.0	2.32	0.0
TAIWAN	713	226.9	35.7	0.0	55.68
SOUTH KOREA	732	104.8	21.0	0.0	64.10
JAPAN	740	861.0	19.7	0.0	4.48
INDIA	750	101.1	9.1	2.10	10.05
PAKISTAN	770	108.5	14.1	0.91	22.94
BURMA	775	71.2	26.9	0.57	2.75
CEYLON	780	144.4	44.3	2.67	6.62
THAILAND	800	128.5	34.4	0.0	9.09
CAMBODIA	811	135.7	25.1	3.43	22.24
LAOS	812	86.5	19.7	1.80	142.05
SOUTH VIETNAM	817	149.7	16.3	0.0	96.40
MALAYSIA	820	305.8	64.9	0.0	3.80
PHILIPPINES	840	159.9	32.1	0.0	13.00
INDONESIA	850	99.2	18.2	3.53	3.91
AUSTRALIA	900	2001.7	14.4	0.0	0.13
NEW ZEALAND	920	1979.9	39.4	0.0	0.11

Here again the unavailability of data made the demonstration artificial. As was asserted in the preceding section, the blocking of countries by external economic ties should depend primarily on pair-wise data. But this researcher did not have access to any standardized, comprehensive pair-wise data. Variables 3 and 4 above are pair-wise but not comprehensive. Foreign aid is provided in substantial amounts by countries other than the United States and the Soviet Union. But this researcher could not locate any but the most piecemeal data on other donors. Variable 2 above is not pair-wise at all. Pair-wise trade data is collected by the International Monetary Fund, and their data is both standardized and reasonably comprehensive. But that data is not made available to the public in comprehensive form.[1] Without pair-wise trade data it is virtually impossible to construct a logical theory for blocking countries by external economic ties. Nevertheless, this demonstration was carried through to completion because its purpose is not to deduce substantive

---

[1] After completion of the research described here, the author did obtain access to the IMF data and began a Cluster Analysis on it. But the results were not obtained in time to incorporate them in this thesis. See Section II.B.7 for a description of the work in progress. The data was obtained through the Inter-University Consortium for Political Research, on computer tape. The reason why the data is not generally available was obvious: its sheer magnitude. For purposes of data collection, the IMF defines 207 countries, and 207 countries taken two at a time produce 21,321 trading combinations. The complete data file contains almost 500,000 numbers.

results but to demonstrate a procedure. And that procedure can still be demonstrated using the foreign aid data (Variables 3 and 4) which is pair-wise, although incomplete.

4. Choice of Dissimilarity Coefficient

This section describes the process of converting the data in Table V to a matrix of Dissimilarity Coefficients.

When Cluster Analysis was used to highlight the interactions of variables (Section II.A.4 above), the choice of DC was motivated by a desire to have all five variables weighted equally, to prevent the user's preconceptions from affecting the results. Precisely the opposite is true here. Here the author presumed that he already knew how the variables interact. He wanted to incorporate that knowledge into the DC. The DC was constructed using the following rationale:

First of all, it was decided that the DC between a foreign aid donor and any other country should be inversely related to the level of that foreign aid. Thus, for a first cut, the formulas

$$DC(US,i) \sim \frac{1}{1 + usaid_i} \quad \text{and} \quad DC(SOV,i) \sim \frac{1}{1 + sovaid_i}$$

were considered. Next it was observed that 27 of the 85 countries received foreign aid from both the United States and the Soviet Union. To incorporate relative dependency into the formulas, it was decided to insert a ratio of aid levels. Hence the following formulas were considered.

$$DC(US,i) \sim \frac{1 + sovaid_i}{1 + usaid_i} \quad \text{and} \quad DC(SOV,i) \sim \frac{1 + usaid_i}{1 + sovaid_i}$$

These formulas seemed reasonable except that the same level of foreign aid has smaller impact on a rich country than it does on a poor country. So it was decided to insert GNP_i as a scaling factor wherever an aid term appeared in either formula. But this insertion tended to greatly reduce the size of the aid terms with respect to the "1" terms. Therefore the "1" terms were arbitrarily reduced to "0.1", producing

$$DC(US,i) = \frac{0.1 + \frac{sovaid_i}{GNP_i}}{0.1 + \frac{usaid_i}{GNP_i}} \quad \text{and} \quad DC(SOV,i) = \frac{0.1 + \frac{usaid_i}{GNP_i}}{0.1 + \frac{sovaid_i}{GNP_i}}$$

The fact that DC(US,i) and DC(SOV,i) are reciprocals and the fact that they are dimensionless had intuitive appeal. The only apparent shortcomings were the two imposed by unavailability of data: aid from other countries is ignored, and pair-wise trade is ignored. Although a total trade figure was available, there seemed to be no logical way to substitute it for the missing pair-wise figure.

At this point it was verified that the DC(US,i) formula would apply to every country dyad in which the United States is a member, except for the United States - Soviet Union dyad. Similarly, the DC(SOV,i) formula applies to every

country dyad in which the Soviet Union is a member, except for the United States - Soviet Union dyad. Thus it remained to construct formulas for the United States - Soviet Union dyad and for all dyads in which neither the United States nor the Soviet Union is a member. Hopefully the same formula would apply to both.

But here the lack of pair-wise trade data was really crippling. The only pair-wise economic ties of any significance involved pair-wise trade. The only logical formula necessarily involved the inverse of pair-wise trade. There seemed no natural way to use the available data on total trade. Finally, in desperation it was rationalized that a country whose foreign trade is large with respect to its GNP tends to have closer ties with another country in the same situation. It was decided that the nucleus of the formula should be

$$DC(i,j) \sim |trade_i - trade_j|$$

But because of the weak theory here as compared to the rigorous formulas for DC(US,i) and DC(SOV,i) it was decided to diminish the effect of trade difference when foreign aid recipients are involved. Hence it was decided to expand the formula to

$$DC(i,j) = \sqrt{|trade_i - trade_j| \frac{sovaid_i + usaid_i + sovaid_j + usaid_j}{4}}$$

The dearth of theoretical foundation here is admitted. It casts suspicion on the values in the DC matrix and on the dendrogram finally obtained. But the reader is again reminded that the purpose here is to demonstrate the procedure, not to deduce substantive results.

Turning from the substance to the procedure, there is a significant departure from normal Cluster Analysis procedure, taken above, that warrants explanation. Two completely different DC formulas have been developed, one to be used when the United States or Soviet Union is a dyad member and another to be used the rest of the time. The mathematical significance of this duality is that the DC formulas, taken collectively, produce gross violations of the metric inequality, which is

$$DC(a,b) + DC(b,c) \geq DC(a,c) \quad \text{for all } a,b,c$$

Generally, it is desirable although not essential that a matrix of DC's satisfy the metric inequality. When they do not, the clustering algorithm can be expected to produce high "distortion" between the matrix of DC's and the dendrogram. (Loosely defined, "distortion" is the difference between DC(i,j) and the level at which country i and country j cluster together in an agglomerative algorithm.) But of what significance is high distortion? The word carries derogatory connotations, but is distortion really undesirable in Cluster Analysis? This author maintains that it depends on the purpose of the clustering. In the "highlighting of variables" application, distortion was not desirable: figuratively speaking,

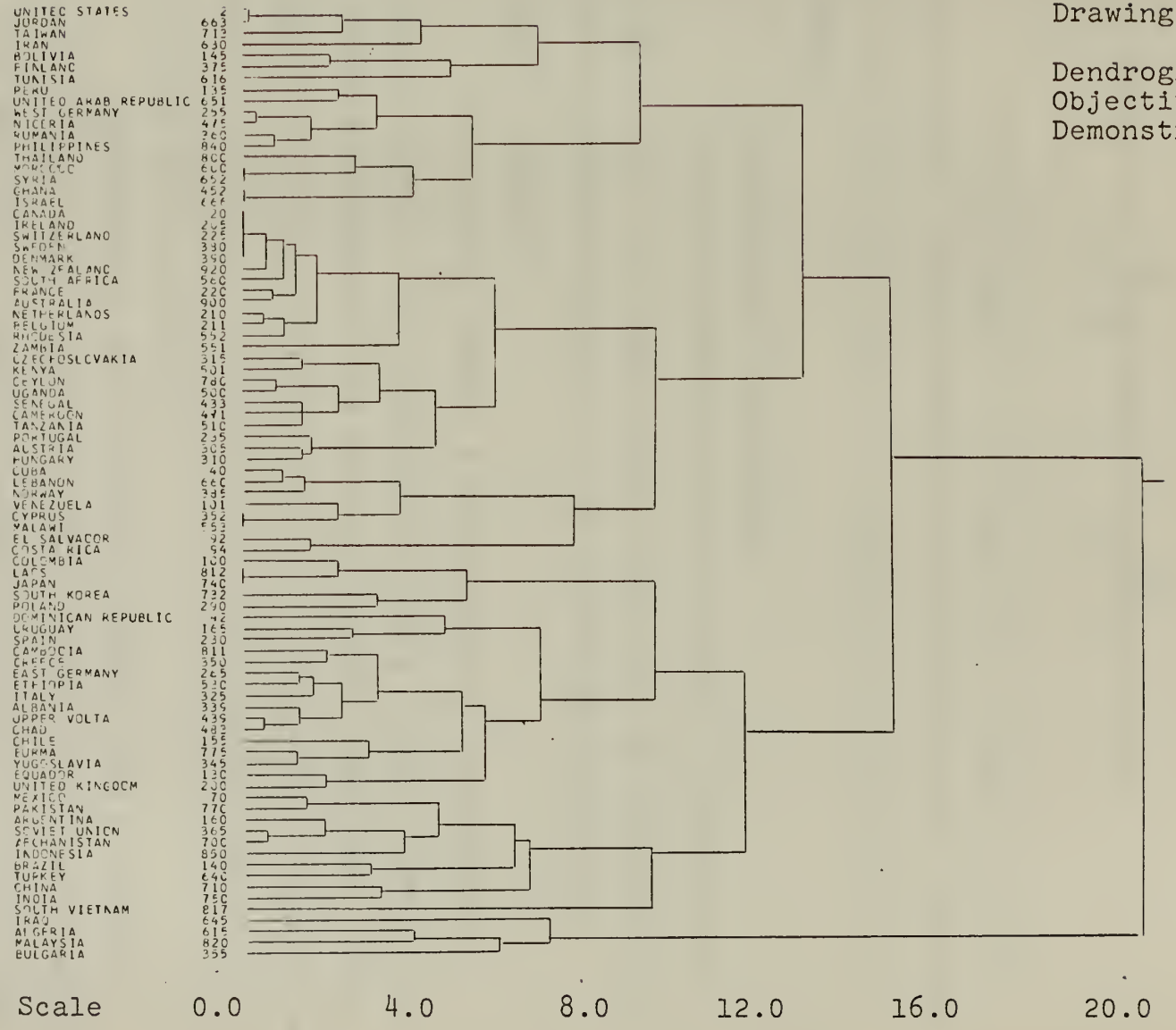
each country had been plotted in five-dimensional space and the clustering algorithm was searching for natural clusters, as plotted. But in this "objective classifying of countries" application, distortion is natural: the original pair-wise similarities specified in the DC matrix cannot be expected to be representable in Euclidean space, and during the clustering it is desired that these original similarities be affected by intermediate countries. With this reasoning, it is asserted that violation of the metric inequality is necessary and that the use of two or more DC formulas is acceptable.

Using the formulas developed above, the 340 data points in Table V were transformed to a matrix of Dissimilarity Coefficients.

5. Choice of Algorithm

Here, as in the highlighting demonstration, the first decision point was to choose among the agglomerative, divisive and reallocative algorithms. Again the divisive algorithms were discarded because they are not as well documented as their agglomerative counterparts. The reallocative algorithms did not apply because they require that the DC be a metric. Hence the agglomerative algorithm was selected.

The final decision was to select a sorting strategy. The Complete Linkage-Furthest Neighbor strategy was eliminated from consideration here; it does not permit any chaining, which is desirable in this application. On the other end of the spectrum, the Single Linkage-Nearest Neighbor sorting strategy maximizes chaining, often to the extent that natural



Drawing 2

Dendrogram Output from Objective Classification Demonstration

clusters are obscured. For this application it was determined to use one of the compromise sorting strategies. Among these, Group Average sorting seemed to correspond with the concepts of bloc membership, ratios of foreign aid, etc. It is expressed mathematically as

$$DC(k,ij) = \frac{n_i}{n_i + n_j} DC(k,i) + \frac{n_j}{n_i + n_j} DC(k,j)$$

The dendrogram in Drawing 2 was obtained using the Group Average sorting strategy in an agglomerative algorithm. But once again the reader is cautioned that the results are suspect.

6. Explanation of Dendrogram

Despite the admitted artificiality of the results obtained here, the layman might appreciate an explanation of the information available in any dendrogram produced by Cluster Analysis.

The key to reading a dendrogram is the concept of "cluster level." By merely specifying a cluster level, the following information can be read from the dendrogram: the number of clusters and the countries contained in each cluster. That is, there is a correspondence from cluster level to a partition of the countries.

The scale at the bottom of Drawing 2 is a cluster level scale. Note that the minimum value of cluster level is 0.0 at the far left and the maximum value is 20.0 at the far

right. A low cluster level specifies a partition having many small clusters, while a high cluster level specifies a partition having a few large clusters. Thus cluster level can be thought of as a measure of the largest dissimilarity (or, equivalently, the weakest bond) present within any cluster in the partition.

For example, consider cluster level 0.0, the minimum observed cluster level in Drawing 2. At cluster level 0.0, the 85 countries are partitioned into 77 clusters. Seventy-two of these 77 contain only a single country. Four of the 77 contain exactly two countries. And one cluster contains 5 countries: Canada, Ireland, Switzerland, Sweden and Denmark. Since 0.0 is the minimum observed cluster level, we may conclude that the strongest possible bonds exist within every cluster. Specifically, we may conclude that Canada, Ireland, Switzerland, Sweden and Denmark are bound together by the tightest possible economic ties. Our mathematical model will not separate them even at the lowest cluster level.

Consider next a slightly higher cluster level, say 1.3. Here we are permitting slightly weaker bonds to be present within clusters. We find that the 85 countries are here partitioned into 55 clusters. Thirty-three of those clusters contain a single country, twelve contain exactly two countries, five contain exactly three countries, one contains four countries, and one contains nine countries. In the nine-country cluster, Canada, Ireland, Switzerland, Sweden and Denmark have been joined by New Zealand, South Africa, France and Australia.

We may conclude that slightly weaker economic ties bind the four new countries to the original five.

Similar inferences can be drawn from any dendrogram produced by Cluster Analysis.

7. Work in Progress

Throughout this second demonstration of Cluster Analysis it has been emphasized that the unavailability of pair-wise trade data made the demonstration artificial. But this artificiality can soon be removed. Pair-wise trade data was recently provided to this author through the Inter-University Consortium for Political Research [Ref. 9]. Time will not permit this author to complete a Cluster Analysis on the data, but if another researcher chooses to undertake it, the following plan of attack is suggested.

a. Step 1 - Reduce data file to manageable size

The ICPR data file contains approximately 333,720 pairwise trade data: annual trade values, in millions of U.S. dollars, for the years 1958 through 1968, among 207 different "countries." Many of these "countries" are actually colonies and many others have negligible foreign trade except with a single "sponsor country." The logical first step is to selectively reduce the size of the data file by eliminating the insignificant "countries", and by selecting a single year and eliminating the other nine. It is recommended that all "countries" be eliminated except the 136 nations having a population of one million or more and those smaller nations having membership in the United Nations as of 1968. These

136 nations are listed on pages 1 through 4 of Ref. 8. It is further recommended that the year 1967 be used and the rest be eliminated temporarily. The author has determined that, through the first one-sixth of the file, 1967 has fewer zero entries than any other year (a zero entry signifies either trade less than 100,000 dollars or missing data). This selective reduction of the data should reduce the file length to about one eighth its original length.

b. Step 2 - Sort and combine data

Preparatory to sorting the data, the reduced data file should be stored on either a disk or a data cell rather than magnetic tape. The ICPR normally provides the data on tape, and tape is a satisfactory input to the data reduction process in step 1 because that process can be sequential, reading the file once from beginning to end. However, the sorting process about to be described cannot read the file sequentially, and magnetic tape is a very inefficient input to processes that must search the data.

The ICPR data file does not list one trade figure per country dyad per year. It lists up to four figures, namely,

1. Value of exports from i to j, as reported by i
2. Value of exports from i to j, as reported by j
3. Value of exports from j to i, as reported by j
4. Value of exports from j to i, as reported by i

Hopefully numbers 1 and 2 are approximately equal and numbers 3 and 4 are approximately equal. If so, then total trade

between i and j is the sum of 1 and 3. It is recommended that this approximate equality be assumed for the initial run of this "sort and combine" process. Then the process is simple: search the file for the first record involving the i-j dyad; identify it with respect to direction of trade, regardless of reporting country; continue searching for the second record involving the i-j dyad; identify it with respect to direction; if the directions are opposite then sum the two values and store them; if the directions are the same then ignore the second value and continue searching for the third record; and so on. The reason why shortcuts are in order for the initial run is that this "sort and combine" process will have to be performed 9180 times (136 countries, taken two at a time, yields 9180 different combinations).

c. Step 3 - Choose a Dissimilarity Coefficient

The following formula is recommended as a DC, at least initially:

$$DC(i,j) = \frac{1}{1 + Trade_{ij}}$$

More elaborate formulas can be developed later by incorporating the rationale in Section II.B.4 of this thesis.

d. Step 4 - Choose a Clustering Algorithm

It is recommended that an agglomerative algorithm with Group Average sorting strategy be used, for the same reasons that it was selected in Section II.B.5 above. This involves making the following additions and substitutions in

the computer program listed at the end of this thesis:

Immediately before DO 73 E=1,N insert the following two statements:

```
RATA = S(A,A)/(S(A,A)+S(B,B))
RATB = S(B,B)/(S(A,A)+S(B,B))
```

In place of

```
DS(E) = AMAX1(S(E,A),S(E,B))
```

substitute

```
DS(E) = RATA*S(E,A) + RATB*S(E,B)
```

Similarly, in place of

```
70 DS(E) = AMAX1(S(A,E),S(B,E))
```

substitute

```
70 DS(E) = RATA*S(A,E) + RATB*S(B,E)
```

And finally, in place of

```
71 DS(E) = AMAX1(S(E,A),S(B,E))
```

substitute

```
71 DS(E) = RATA*S(E,A) + RATB*S(B,E)
```

III. CONCLUSION

This thesis has demonstrated two potential uses of Cluster Analysis in which the nations of the world are treated as measurable objects. The substantive results obtained in each demonstration are not presented as conclusions; they were derived incidentally while demonstrating methods. It is asserted that the two uses illustrated here, markedly different in several respects, are representative of a wide range of applications for Cluster Analysis in the fields of political science and international relations. Although Cluster Analysis was developed for the physical sciences and has so far received scant attention outside that context, it is readily adaptable to the social sciences. In particular, it is extremely well suited to model building and statistical analysis involving the nations of the world. As such, it warrants the attention of the U.S. State Department.

APPENDIX A

DATA

Except for three data points, all data used in this thesis were made available by the Inter-University Consortium for Political Research. The data were originally collected by Charles Lewis Taylor and Michael C. Hudson. Neither the original collectors of the data nor the consortium bear any responsibility for the analysis or interpretations presented here.

Following are the precise definitions of the nine variables used in this thesis. All definitions are extracted verbatim from Ref. 8.

Variable name: Concentration of Population in Cities, 1965
Definition: Concentration is defined as: the sum over all cities of the squares of the proportion of the total population residing in each city. Concentration is higher the fewer cities and the greater the size of the largest city relative to the total population. [Ref. 8, p. 16]

Variable name: Radios per 1000 Population, 1965
Definition: Figures relate to all types of receivers including those connected to a re-distribution system. They relate either to the number of licenses issued or sets declared or to the estimated number of receivers in use. In many countries a license may cover more than one receiver in the same household. Data exclude television sets. [Ref. 8, p. 32]

Variable name: Students in Higher Education (Third Level) per One Million Population, 1965.
Definition: Data refer to the enrollment in all institutions of education at the third level, i.e., degree granting and non-degree granting institutions of both private and public higher education of all types. These include universities, higher technical schools, teacher training schools, theological schools, etc. As far as possible part time students are included in the figures but correspondence courses and auditors are generally excluded. [Ref. 8, p. 41]

Variable name: Ethno-Linguistic Fractionalization
Definition: The main source for this variable (Atlas Narodov Mira) makes little distinction between ethnic and linguistic differences in its definition and collection of data. Groups are determined not by their physical characteristics but by their roles, their descents and their relationships to others. An index of fractionalization calculated upon data from Atlas does correlate highly with a similar index calculated upon linguistic data from other sources, but not quite highly enough to be considered the same indicator. Other sources used here report only linguistic data. Index of fractionalization was calculated by the following formula:

$$F = 1 (N \text{ subi} / N) (N \text{ subi} - 1/N-1)$$

where N subi = number of people in the ith group
and N = total population [Ref. 8, p. 46]

Variable name: Press Freedom Index, 1965

Definition: This index, created by the School of Journalism, University of Missouri, is "designed to measure the independence of a nation's broadcasting and press system and its ability to criticize its own local and national governments." The index is comprised of the judgements of panels of native and foreign newsmen on 23 aspects of the press (e.g., extent of legal controls, licensing, government ownership, criticism and censorship). For a fuller description, see Ralph L. Lowenstein, "PICA (Press Independence and Critical Ability) Index: Measuring World Press Freedom," University of Missouri, School of Journalism Freedom of Information Center Publication #166 (August, 1966). The index, which consists of averages of the judges' scores, has a range from -4.00 for less freedom to +4.00 for more. [Ref. 8, p. 116]

Variable name: Gross National Product per Capita, 1965

Definition: This variable was derived by dividing Gross National Product in millions of U.S. dollars by total population in thousands. Gross National Product is reported in constant U.S. dollars and refers to gross national product even for countries which normally report their national accounts in terms of net material product or other concepts. [Ref. 8, p. 65]

Variable name: Trade as percentage of Gross National Product, 1965.

Definition: This variable was derived by dividing total trade

(imports plus exports, merchandise only) by Gross National Product. [Ref. 8, p. 69]

Variable name: Soviet Aid per Capita, 1954 - 1965
Definition: This variable was derived by dividing total Soviet aid by total population. Total Soviet aid data refer to Soviet economic credits and grants to countries in terms of thousand U.S. dollars for the period 1954/5 - 1965. [Ref. 8, p. 107]

Variable name: U.S. Economic Aid per Capita, 1958 - 1965
Definition: This variable was derived by dividing total U.S. economic aid by total population. Total U.S. economic aid data refer to grants and loans and are given in millions of U.S. dollars for the period July 1, 1958 through June 30, 1965. [Ref. 8, p. 107]

The three data points not provided by the ICPR are listed below. The ICPR data file listed all three as missing data. But in each case this author preferred to introduce an approximate (or even erroneous) value rather than eliminate the particular country from the Cluster Analysis. Hence the three values were estimated in the manner specified. Note that no two estimations involved the same country. All countries missing two or more data (among the nine variables used) in the ICPR data file were omitted from the Cluster Analysis at the outset.

Country: Chile

Variable name: Radios per 1000 Population, 1965

Estimated value: 240.0

Method of estimation: Average of values for Peru and Argentina.

Country: Chad

Variable: Students in Higher Education (Third Level) per One Million Population, 1965

Estimated value: 230.0

Method of estimation: Average of values for Mali, Upper Volta, Sudan and Cameroon.

Country: Zambia

Variable: Students in Higher Education (Third Level) per One Million Population, 1965

Estimated value: 170.0

Method of estimation: Average of seventeen neighboring countries.

COMPUTER PROGRAM

Clustering by Agglomerative Algorithm with Complete Link Sorting Strategy

```
      IMPLICIT INTEGER(A,B,E)
      DIMENSION DS(85)
      COMMON N, NOLD, A, B, E, H, S(85,86), NTIES
      N=85
      DO 51 I=1,N
      DO 50 J=I,N
      S(I,J)=0.0
   50 CONTINUE
   51 CONTINUE
      DO 52 I=1,N
      S(I,I)=100.0
   52 CONTINUE
      DO 49 I=1,N
      READ (5,1000) NACODE
 1000 FORMAT (21X,I3)
      S(I,86) = NACODE
   49 CONTINUE
   53 READ (5,1001,END=54) I,J,S(I,J),I,J,S(I,J),I,J,S(I,J),I,J,S(I,J),
     1I,J,S(I,J)
 1001 FORMAT (5(I2,I3,F9.4,1X))
      GO TO 53
C
CCCCC
CCCCC                                  ADDRESS 54 IS DECISION POINT
C                                      FOR STOPPING, PRINTING
                                       MATRIX, PUNCHING MATRIX, ETC.
   54 IF (N.EQ.85) CALL RITMAT(6)                                  OPTION
      IF (N.EQ.9) CALL RITMAT(6)                                   OPTION
      IF (N.GT.1) GO TO 58
      CALL RITMAT(6)
      DO 55 JA=1,85
      J = 87-JA
      NACODE = S(1,J)
      WRITE (7,1005) NACODE, JA
 1005 FORMAT (I3,73X,I2)
   55 CONTINUE
      STOP
C
CCCC
C
```

```
C
CC                                                    SEARCH MATRIX FOR MIN
   58 A=1
      B=2
      H=S(2,1)
      DO 63 I=3,N
      IM1=I-1
      DO 62 J=1,IM1
      IF (S(I,J)-H) 60,61,62
   60 H=S(I,J)
      B=I
      A=J
      NTIES=0
      GO TO 62
   61 NTIES=NTIES+1
   62 CONTINUE
   63 CONTINUE
      CALL RITMIN(6)
CCCCCCCC
                                                      COMPUTE DS VECTOR, USING
                                                      COMPLETE LINK METHOD
CC
      DO 73 E=1,N
      IF (E.LE.A) GO TO 70
      IF (E.LE.B) GO TO 71
CCCC                                                  A = COLUMN #
                                                      B = ROW #
CCCC                                                  A < B IS ALWAYS TRUE IN
                                                      SOUTHWEST HALF OF MATRIX
      DS(E) = AMAX1(S(E,A),S(E,B))                                    OPTION
      GO TO 73
   70 DS(E) = AMAX1(S(A,E),S(B,E))                                    OPTION
      GO TO 73
   71 DS(E) = AMAX1(S(E,A),S(B,E))                                    OPTION
   73 CONTINUE
CCCC
C
```

```
C
C                                                    PLUG DS INTO S
      DO 82 E=1,N
      IF (E-A) 80,82,81
   80 S(A,E)=DS(E)
      GO TO 82
   81 S(E,A)=DS(E)
   82 CONTINUE
      S(A,A)=S(A,A)+S(B,B)
      IF (B.EQ.N) GO TO 89
      DO 84 E=1,N
      DS(E) = S(N,E)
   84 CONTINUE
      NM1 = N-1
      DO 88 E=1,NM1
      IF (E-B) 85,86,87
   85 S(B,E) = DS(E)
      GO TO 88
   86 S(B,B) = DS(N)
      GO TO 88
   87 S(E,B) = DS(E)
   88 CONTINUE
C
C                                                    SHUFFLE COUNTRY CODES
   89 AP1 = A+1
      DO 91 J=AP1,86
      IF (S(A,J).GT.0.1) GO TO 92
   91 CONTINUE
   92 JA = J-1
      J=86
   93 IF (S(B,J).LT.0.1.OR.J.EQ.B) GO TO 94
      S(A,JA) = S(B,J)
      S(B,J) = 0.0
      JA = JA-1
      J = J-1
      GO TO 93
   94 NP1 = N+1
      DO 95 J=NP1,86
      S(B,J) = S(N,J)
   95 CONTINUE
C
      NOLD=N
      N=N-1
      CALL RITCLU(6)
      GO TO 54
      END
```

```
C
      SUBROUTINE RITMAT(NPRINT)
C
      COMMON N, NOLD, A, B, E, H, S(85,86), NTIES
      WRITE (NPRINT,1025) N
 1025 FORMAT ('1N=',I3)
      DO 24 JSTART=1,71,14
      JSTOP = JSTART + 13
      DO 22 I=1,N
      WRITE (NPRINT,1026) I, (S(I,J),J=JSTART,JSTOP)
 1026 FORMAT (' ',I3,1X,14F9.4)
   22 CONTINUE
      WRITE (NPRINT,1027) (J,J=JSTART,JSTOP)
 1027 FORMAT (' ',14I9,//)
   24 CONTINUE
      DO 28 I=1,N
      WRITE (NPRINT,1028) I, S(I,85), S(I,86)
 1028 FORMAT (' ',108X,I3,1X,F9.4,F9.4)
   28 CONTINUE
      WRITE (NPRINT,1029)
 1029 FORMAT (' ',115X,'85       86',//)
      RETURN
      END
```

```
C
      SUBROUTINE RITMIN(NPRINT)
C
      IMPLICIT INTEGER(A,B,E)
      COMMON N, NOLD, A, B, E, H, S(85,86), NTIES
      WRITE (NPRINT,1009) N
      WRITE (NPRINT,1010) H
      WRITE (NPRINT,1011) B,A
      WRITE (NPRINT,1012) NTIES
 1009 FORMAT ('ON=',I3)
 1010 FORMAT (' H = SMALLEST MATRIX ELEMENT =',F9.4)
 1011 FORMAT (' H LOCATION IS   S(',I3,I3,')')
 1012 FORMAT (' NTIES =',I3,//)
      RETURN
      END
```



```
C
      SUBROUTINE RITCLU(NPRINT)
C
      IMPLICIT INTEGER(A,B,E)
      COMMON N, NOLD, A, B, E, H, S(85,86), NTIES
      WRITE (NPRINT,1030) H,B,A,NOLD,B,NOLD,N
 1030 FORMAT ('OAT THE',F9.4,' LEVEL, ROW',I3,' CLUSTERS INTO ROW',I3,/,
     1' ',23X,'ROW',I3,' MOVES UP INTO ROW',I3,/,
     2' AND N DECREMENTS FROM',I3,' TO',I3,//)
      RETURN
      END
```

LIST OF REFERENCES

1. Ball, G. H. and Hall, D. H., "A Clustering Technique for Summarizing Multivariate Data," Behavioral Science, v. 12, p. 153-155, 1967.

2. Redecker, J. B., Planning as an Instrument for Decision-making in Foreign Affairs: Diagnosis and System Design, MS Thesis, Massachusetts Institute of Technology, 1970.

3. Wilhelm, J. K., Foreign Policy Objectives: A Systematic Approach to Management, Evaluation, Coordination and Resource Allocation, paper presented at the 25th Military Operations Research Society, Intelligence Working Group, New London, Connecticut, 17 June 1970.

4. Cormack, R. M., "A Review of Classification," Journal of the Royal Statistical Society, v. 134, p. 321-353, Part 3, 1971.

5. Sneath, P. H. A. and Sokal, R. R., Principles of Numerical Taxonomy, Freeman, 1963.

6. Jardine, N. and Sibson, R., Mathematical Taxonomy, Wiley 1971.

7. Carlson, P. E. and Hancock, W. J., Law of the Sea: An Application of Cluster Analysis, MS Thesis, Naval Post-graduate School, 1972.

8. Taylor, C. L., and Hudson, M. C., World Handbook of Political and Social Indicators II, Section I, Cross-National Aggregate Data, Yale University, 1970.

9. Gillespie, J., and Zinnes, D., World Trade Data: 1958 - 1968, Inter-University Consortium for Political Research, 1970.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst. Professor B. O. Shubert Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
4.	Asst. Professor E. J. Laurance Department of Government and Humanities Naval Postgraduate School Monterey, California 93940	1
5.	LCDR James Richard Lamping, USN 15 La Playa Monterey, California 93940	1
6.	Professor R. von Pagenhardt Naval Management Systems Center Naval Postgraduate School Monterey, California 93940	1
7.	Mr. Dan Daniels M/MS Room 2231 Department of State Washington, D. C. 20520	1
8.	Inter-University Consortium for Political Research Box 1248 Ann Arbor, Michigan 48106	1
9.	Naval Postgraduate School Department of Operations Research and Administrative Sciences Monterey, California 93940	1
10.	Chief of Naval Personnel Pers 11b Department of the Navy Washington, D. C. 20370	1

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author)
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

REPORT TITLE
Applications of Cluster Analysis to Some Problems of Interest to the U. S. Department of State

DESCRIPTIVE NOTES (Type of report and inclusive dates)
Master's Thesis; September 1973

AUTHOR(S) (First name, middle initial, last name)
James Richard Lamping

REPORT DATE
September, 1973

7a. TOTAL NO. OF PAGES
60

7b. NO. OF REFS
9

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Naval Postgraduate School
Monterey, California 93940

13. ABSTRACT

This thesis asserts that Cluster Analysis, or Numerical Taxonomy, has many potential applications in the field of international relations. It demonstrates two representative applications. Both examples treat the nations of the world as objects having measurable attributes, and both examples use selected attributes to produce a dendrogram (or hierarchical classification) of the nations of the world. In one example this dendrogram is used to objectively group the nations into blocs based on external economic ties. In the other example the dendrogram is used to highlight interactions among five attributes, ignoring the identity of individual nations, the same way a scatter plot highlights interactions between two variables.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Cluster Analysis						
Dissimilarity Coefficient						
Dendrogram						

Thesis 145424
L2556 Lamping
c.1 Applications of cluster analysis to some problems of interest to the U. S. Department of State.